

The genetic architecture of target-site resistance to pyrethroid insecticides in the African malaria vectors *Anopheles gambiae* and *Anopheles coluzzii*

Chris S. Clarkson^{1,*}, Alistair Miles^{2,1,*}, Nicholas J. Harding², David Weetman³, Dominic Kwiatkowski^{1,2}, Martin Donnelly^{3,1}, and The *Anopheles gambiae* 1000 Genomes Consortium⁴

¹Wellcome Sanger Institute, Hinxton, Cambridge CB10 1SA

²Big Data Institute, University of Oxford, Li Ka Shing Centre for Health Information and Discovery, Old Road Campus, Oxford OX3 7LF

³Liverpool School of Tropical Medicine, Pembroke Place, Liverpool L3 5QA

⁴<https://www.malariagen.net/projects/ag1000g#people>

*These authors contributed equally

14th November 2019

Abstract

Resistance to pyrethroid insecticides is a major concern for malaria vector control, because these are the compounds used in almost all insecticide-treated bed-nets (ITNs), and are also widely used for indoor residual spraying (IRS). Pyrethroids target the voltage-gated sodium channel (VGSC), an essential component of the mosquito nervous system, but substitutions in the amino acid sequence can disrupt the activity of these insecticides, inducing a resistance phenotype. Here we use Illumina whole-genome sequence data from phase 2 of the *Anopheles gambiae* 1000 Genomes Project (Ag1000G) to provide a comprehensive account of genetic variation in the *Vgsc* gene in mosquito populations from 13 African countries. In addition to the three known *kdr*

resistance alleles, we describe 20 non-synonymous nucleotide substitutions at appreciable frequency in one or more populations that are previously unknown in *Anopheles* mosquitoes. Thirteen of these novel alleles were found to occur almost exclusively on haplotypes carrying the known L995F resistance allele (L1014F in *Musca domestica* codon numbering), and may enhance or compensate for the L995F resistance phenotype. A novel mutation I1527T, which is adjacent to a predicted pyrethroid binding site, was found in tight linkage with either of two alleles causing a V402L substitution, similar to a combination of substitutions found to cause pyrethroid resistance in several other insect species. We analyse the genetic backgrounds on which non-synonymous alleles are found, to determine which alleles have experienced recent positive selection, and to refine our understanding of the spread of resistance between species and geographical locations. We describe ten distinct *kdr* carrying haplotype groups with evidence of recent positive selection, five of which carry the known L995F resistance allele, five of which carry the known L995S resistance allele. Five of these groups are localised to a single geographical location, and five comprise haplotypes from different countries, in one case separated by over 3000 km, providing new information about the geographical distribution and spread of resistance. Two "non-*kdr*" haplotype groups with evidence of recent selection were also detected, one of which carries the novel I1527T allele, and one of which carries a novel M490I allele. We also find evidence for multiple introgression events transmitting resistance alleles between *An. gambiae* and *An. coluzzii*. We identify markers that could be used to design high-throughput, low-cost genetic assays for improved surveillance of pyrethroid resistance in the field. Our results demonstrate that the molecular basis of target-site pyrethroid resistance in malaria vectors is more complex than previously appreciated, and provide a foundation for the development of new genetic tools to track the spread insecticide resistance and improve the design of strategies for insecticide resistance management.

Introduction

Pyrethroid insecticides have been the cornerstone of malaria prevention in Africa for almost two decades [1]. Pyrethroids are currently used in all insecticide-treated bed-nets (ITNs), and are widely used in indoor residual spraying (IRS) campaigns as well as in agriculture. Pyrethroid resistance is widespread in malaria vector populations across Africa [2]. The World Health Organization (WHO) has published plans for insecticide resistance

management (IRM), which emphasise the need for improvements in both our knowledge of the molecular mechanisms of resistance and our ability to monitor them in natural populations [3, 4].

The voltage-gated sodium channel (VGSC) is the physiological target of pyrethroid insecticides, and is integral to the insect nervous system. Pyrethroid molecules bind to sites within the protein channel and prevent normal nervous system function, causing paralysis (“knock-down”) and then death. However, amino acid substitutions at key positions within the protein alter the interaction with insecticide molecules (target-site resistance), increasing the dose of insecticide required for knock-down (hence this type of resistance is also known as knock-down resistance or *kdr* [5, 6]. In the African malaria vectors *Anopheles gambiae* and *An. coluzzii*, three substitutions have been found to cause pyrethroid resistance. Two of these substitutions occur in codon 995¹, with L995F prevalent in West and Central Africa [7, 8], and L995S found in Central and East Africa [9, 8]. A third substitution, N1570Y, has been found in West and Central Africa and shown to increase resistance in association with L995F [11]. However, studies in other insect species have found a variety of other *Vgsc* substitutions inducing a resistance phenotype [12, 13, 6]. To our knowledge, no studies in malaria vectors have analysed the full *Vgsc* coding sequence, thus the molecular basis of target-site resistance to pyrethroids has not been fully explored.

Basic information is also lacking about the spread of pyrethroid resistance in malaria vectors [3]. For example, it is not clear when, where or how many times pyrethroid target-site resistance has emerged. Geographical paths of transmission, carrying resistance alleles between mosquito populations, are also not known. Previous studies have found evidence that L995F occurs on several different genetic backgrounds, suggesting multiple independent outbreaks of resistance driven by this allele [14, 15, 16, 17]. However, these studies analysed only small gene regions in a limited number of mosquito populations, and therefore had limited resolution to make inferences about relationships between haplotypes carrying this allele. It has also been shown that the L995F allele spread from *An. gambiae* to *An. coluzzii* in West Africa [18, 19, 20, 21]. However, both L995F and L995S now have

¹Codon numbering is given here relative to transcript AGAP004707-RD as defined in the AgamP4.12 gene-set annotations. A mapping of codon numbers from AGAP004707-RD to *Musca domestica*, the system in which *kdr* mutations were first described [10], is given in Table 1.

85 wide geographical distributions [8], and to our knowledge no attempts have been made to
86 infer or track the geographical spread of either allele across Africa.

87 Here we report an in-depth analysis of genetic variation in the *Vgsc* gene, using whole-
88 genome Illumina sequence data from phase 2 of the *Anopheles gambiae* 1000 Genomes
89 Project (Ag1000G) [22]. The Ag1000G phase 2 resource includes data on nucleotide vari-
90 ation in 1,142 wild-caught mosquitoes sampled from 13 countries, with representation of
91 West, Central, Southern and East Africa, and of both *An. gambiae* and *An. coluzzii*.
92 We investigate variation across the complete gene coding sequence, and report popula-
93 tion genetic data for both known and novel non-synonymous nucleotide substitutions. We
94 then use haplotype data from the chromosomal region spanning the *Vgsc* gene to study
95 the genetic backgrounds carrying resistance alleles, infer the geographical spread of resis-
96 tance between mosquito populations, and provide evidence for recent positive selection.
97 Finally, we explore ways in which variation data from Ag1000G can be used to design
98 high-throughput, low-cost genetic assays for surveillance of pyrethroid resistance, with
99 the capability to differentiate and track resistance outbreaks.

100 Results

101 *Vgsc* non-synonymous nucleotide variation

102 To identify variants with a potentially functional role in pyrethroid resistance, we ex-
103 tracted single nucleotide polymorphisms (SNPs) that alter the amino acid sequence of the
104 VGSC protein from the Ag1000G phase 2 data resource. We then computed their allele
105 frequencies among 16 mosquito populations defined by species and country of origin. Al-
106 leles that confer resistance are expected to increase in frequency under selective pressure,
107 and we filtered the list of potentially functional variant alleles to retain only those at or
108 above 5% frequency in one or more populations (Table 1). The resulting list comprises
109 23 variant alleles, including the known L995F, L995S and N1570Y resistance alleles, and a
110 further 20 alleles not previously described in anopheline mosquitoes. We reported 12 of
111 these novel alleles in our overall analysis of the Ag1000G phase 1 data resource [22], and
112 we extend the analyses here to incorporate SNPs which alter codon 531, 697, 1507, 1603
113 and two tri-allelic SNPs affecting codons 402 and 490.

Table 1. Non-synonymous nucleotide variation in the voltage-gated sodium channel gene. AO=Angola; GH=Ghana; BF=Burkina Faso; CI=Côte d’Ivoire; GN=Guinea; GW=Guinea-Bissau; GM=Gambia; CM=Cameroon; GA=Gabon; UG=Uganda; GQ=Bioko; FR=Mayotte; KE=Kenya; *Ac*=*An. coluzzii*; *Ag*=*An. gambiae*. Species status of specimens from Kenya and Guinea-Bissau is uncertain [22]. All variants are at 5% frequency or above in one or more of the 9 Ag1000G phase 1 populations, with the exception of 2,400,071 G>T which is only found in the CMAg population at 0.4% frequency but is included because another mutation (2,400,071 G>A) is found at the same position causing the same amino acid substitution (M490I); and 2,431,019 T>C (F1920S) which is at 4% frequency in GAAg but also found in CMAg and linked to L995F.

Variant				Population allele frequency (%)																
Position ¹	Ag ²	Md ³	Domain ⁴	AOAc	GHAc	BFAc	CIAc	GNAc	GW	GM	CMAg	GHAg	BFAg	GNAg	GAAg	UGAg	GQAg	FRAg	KE	
2,390,177 G>A	R254K	R261	IL45	0	1	0	0	0	0	0	31	0	0	0	20	0	0	0	0	
2,391,228 G>C	V402L	V410	IS6	0	13	7	8	12	0	0	0	0	0	0	0	0	0	0	0	
2,391,228 G>T	V402L	V410	IS6	0	5	6	0	0	0	0	0	0	0	0	0	0	0	0	0	
2,399,997 G>C	D466H	-	LI/II	0	0	0	0	0	0	0	7	0	0	0	0	0	0	0	0	
2,400,071 G>A	M490I	M508	LI/II	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	19	
2,400,071 G>T	M490I	M508	LI/II	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
2,402,466 G>T	G531V	G549	LI/II	0	0	0	0	0	0	0	0	0	0	0	1	0	6	0	0	
2,407,967 A>C	Q697P	Q724	LI/II	0	0	0	0	0	0	0	0	0	0	0	0	0	6	0	0	
2,416,980 C>T	T791M	T810	IIS1	0	1	2	0	0	0	0	0	29	15	11	0	0	0	0	0	
2,422,651 T>C	L995S	L1014	IIS6	0	0	0	0	0	0	0	16	0	0	0	67	100	0	0	76	
2,422,652 A>T	L995F	L1014	IIS6	84	82	85	92	88	0	0	53	100	100	100	33	0	0	0	0	
2,429,556 G>A	V1507I	-	IIIL56	0	0	0	0	12	0	0	0	0	0	0	0	0	0	0	0	
2,429,617 T>C	I1527T	I1532	IIIS6	0	17	13	8	12	0	0	0	0	0	0	0	0	0	0	0	
2,429,745 A>T	N1570Y	N1575	LIII/IV	0	0	27	0	0	0	0	6	17	21	9	0	0	0	0	0	
2,429,897 A>G	E1597G	E1602	LIII/IV	0	0	0	0	0	0	0	0	0	7	6	0	0	0	0	0	
2,429,915 A>C	K1603T	K1608	IVS1	0	5	5	0	0	0	0	0	0	0	0	0	0	0	0	0	
2,430,424 G>T	A1746S	A1751	IVS5	0	0	0	0	0	0	0	0	29	14	10	0	0	0	0	0	
2,430,817 G>A	V1853I	V1858	COOH	0	0	0	0	0	0	0	0	54	5	6	0	0	0	0	0	
2,430,863 T>C	I1868T	I1873	COOH	0	0	0	0	0	0	0	0	0	26	20	0	0	0	0	0	
2,430,880 C>T	P1874S	P1879	COOH	0	3	21	35	0	0	0	0	0	0	0	0	0	0	0	0	
2,430,881 C>T	P1874L	P1879	COOH	0	0	7	1	25	0	0	0	0	23	48	0	0	0	0	0	
2,431,061 C>T	A1934V	A1939	COOH	0	2	11	46	50	0	0	0	0	0	0	0	0	0	0	0	
2,431,079 T>C	I1940T	I1945	COOH	0	12	4	0	0	0	0	7	0	0	0	0	0	0	0	0	

¹ Position relative to the AgamP3 reference sequence, chromosome arm 2L. Variants marked with an asterisk (*) failed conservative variant filters applied genome-wide in the Ag1000G phase 1 AR3 callset, but appeared sound on manual inspection of read alignments.

² Codon numbering according to *Anopheles gambiae* transcript AGAP004707-RA in geneset AgamP4.4.

³ Codon numbering according to *Musca domestica* EMBL accession X96668 [10].

⁴ Location of the variant within the protein structure. Transmembrane segments are named according to domain number (in Roman numerals) followed by ‘S’ then the number of the segment; e.g., ‘IIS6’ means domain two, transmembrane segment six. Internal linkers between segments within the same domain are named according to domain (in Roman numerals) followed by ‘L’ then the numbers of the linked segments; e.g., ‘IL45’ means domain one, linker between transmembrane segments four and five. Internal linkers between domains are named ‘L’ followed by the linked domains; e.g., ‘LI/II’ means the linker between domains one and two. ‘COOH’ means the internal carboxyl tail.

114 The two known resistance alleles affecting codon 995 had the highest overall allele fre-
115 quencies within the Ag1000G phase 1 cohort (Table 1). The L995F allele was at high
116 frequency in populations of both species from West, Central and Southern Africa . The
117 L995S allele was at high frequency among *An. gambiae* populations from Central and East
118 Africa. Both alleles were present in *An. gambiae* populations sampled from Cameroon
119 and Gabon, including some individuals with a hybrid L995F/S genotype (46/275 individ-
120 uals in Cameroon, 36/56 in Gabon). In Cameroon these alleles were in Hardy Weinberg
121 equilibrium ($\chi^2 = 0.02$, $P > 0.05$), but there was an excess of heterozygotes in Gabon (χ^2
122 $= 8.96$, $P < 0.005$), suggesting a fitness advantage for mosquitoes carrying both alleles at
123 least in some circumstances.

124 The N1570Y allele was present in Guinea, Burkina Faso (both species) and Cameroon.
125 This allele has only ever been found in association with L995F [11], and has been shown
126 to substantially increase pyrethroid resistance when it occurs in combination with L995F,
127 both in association tests of phenotyped field samples [11] and experimentally [23]. To
128 study the patterns of association among non-synonymous variants, we used haplotypes
129 from the Ag1000G phase 1 resource to compute the normalised coefficient of linkage dis-
130 equilibrium (D') between all pairs of variant alleles (Figure 1). As expected, we found
131 N1570Y in almost perfect linkage with L995F, meaning that N1570Y was almost only ever
132 found on haplotypes carrying L995F. Of the 20 novel non-synonymous alleles, 13 also oc-
133 curred almost exclusively in combination with L995F, exhibiting the same LD pattern as
134 N1570Y (Figure 1). These included two variants in codon 1874 (P1874S, P1874L), one
135 of which (P1874S) has previously been associated with pyrethroid resistance in the crop
136 pest *Plutella xylostella* [24]. The abundance of high-frequency non-synonymous variants
137 occurring in combination with L995F is striking for two reasons. First, *Vgsc* is a highly
138 conserved gene, expected to be under strong functional constraint and therefore purify-
139 ing selection, and so any non-synonymous variants are expected to be rare [12]. Second,
140 in contrast with L995F, we did not observe any high-frequency non-synonymous variants
141 occurring in combination with L995S. This contrast was highly significant when data on
142 all variants within the gene were considered: relative to haplotypes carrying the wild-type
143 L995 allele, the ratio of non-synonymous to synonymous nucleotide diversity (π_N/π_S) was
144 28.1 (95% CI [25.2, 31.2]) times higher among haplotypes carrying L995F but 1.5 (95%
145 CI [0.8, 2.2]) times higher among haplotypes carrying L995S. These results indicate that
146 L995F has substantially altered the selective regime for other amino acid positions within

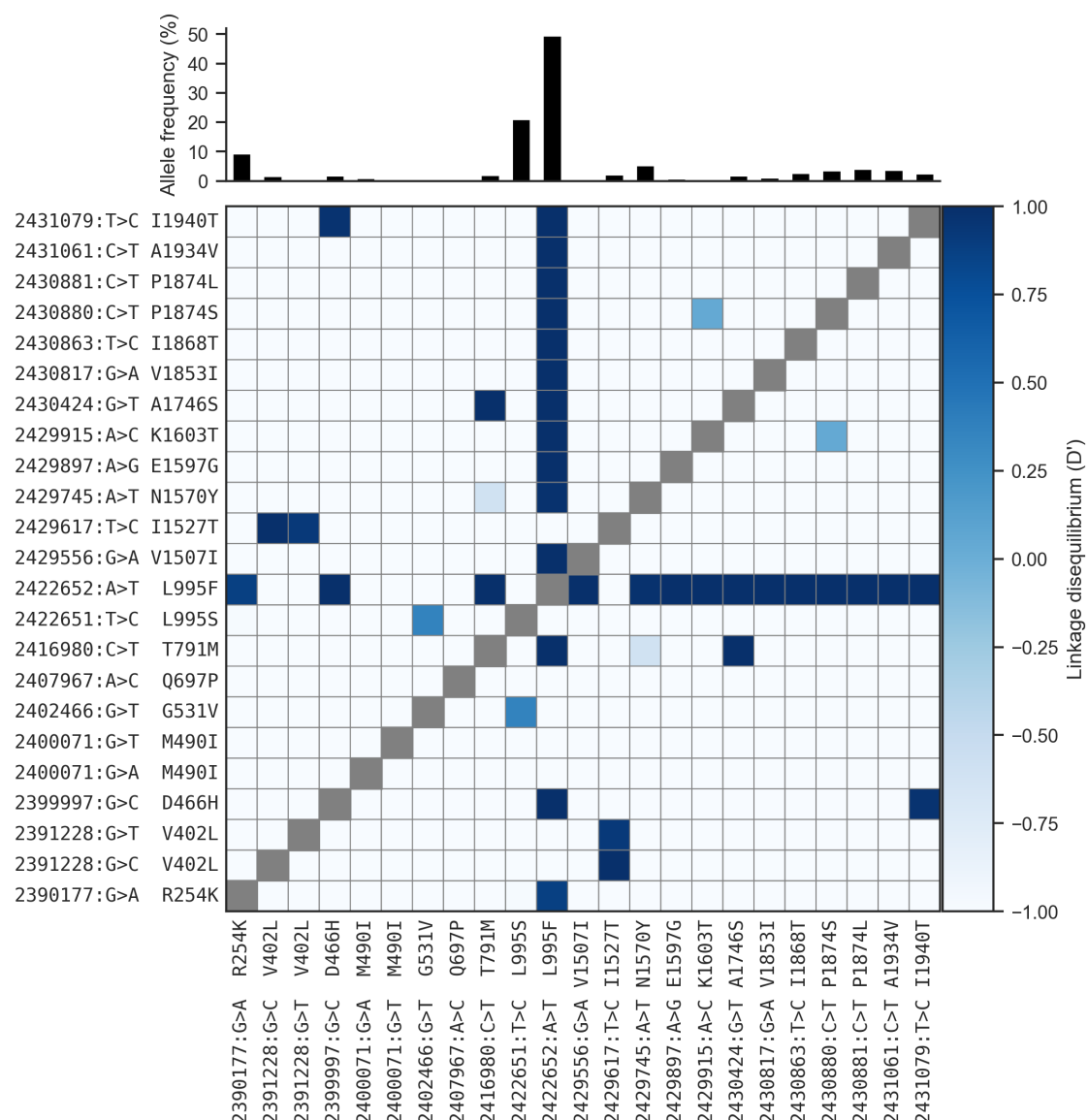


Figure 1. Linkage disequilibrium (D') between non-synonymous variants. A value of 1 indicates that two alleles are in perfect linkage, meaning that one of the alleles is only ever found in combination with the other. Conversely, a value of -1 indicates that two alleles are never found in combination with each other. The bar plot at the top shows the frequency of each allele within the Ag1000G phase 1 cohort. See Table 1 for population allele frequencies.

147 the protein. A number of secondary substitutions have occurred and risen in frequency,
 148 suggesting that they are providing some selective advantage in the presence of insecticide
 149 pressure.

150 A novel allele, I1527T, was present in *An. coluzzii* from Burkina Faso at 14% fre-
 151 quency. Codon 1527 occurs within trans-membrane segment IIIS6, immediately adjacent
 152 to residues within a predicted binding site for pyrethroid molecules, thus it is plausible that

153 I1527T could alter pyrethroid binding [25, 6]. We also found that the two variant alleles
 154 affecting codon 402, both of which induce a V402L substitution, were in strong linkage
 155 with I1527T ($D' \geq 0.8$; Figure 1), and almost all haplotypes carrying I1527T also carried a
 156 V402L substitution. Substitutions in codon 402 have been found in a number of other insect
 157 species and shown experimentally to confer pyrethroid resistance [6]. Because of the lim-
 158 ited geographical distribution of these alleles, we hypothesize that the I1527T+V402L com-
 159 bination represents a pyrethroid resistance allele that arose in West African *An. coluzzii*
 160 populations. However, the L995F allele is at higher frequency (85%) in our Burkina Faso
 161 *An. coluzzii* population, and is known to be increasing in frequency [26], therefore L995F
 162 may provide a stronger resistance phenotype and is replacing I1527T+V402L.

163 The remaining 4 novel alleles (two separate nucleotide substitutions causing M490I;
 164 A1125V; V1254I) did not occur in combination with any known resistance allele (Table 1).
 165 All are private to a single population, and to our knowledge none have previously been
 166 found in other species [13, 6].

167 Genetic backgrounds carrying resistance alleles

168 The Ag1000G data resource provides a rich source of information about the spread of
 169 insecticide resistance alleles in any given gene, because data are available not only for
 170 SNPs in protein coding regions, but also SNPs in introns and flanking intergenic regions,
 171 and in neighbouring genes. These additional variants can be used to analyse the genetic
 172 backgrounds (haplotypes) on which resistance alleles are found. In our initial report of
 173 the Ag1000G phase 1 resource [22], we used 1710 biallelic SNPs from within the 73.5 kbp
 174 *Vgsc* gene (1607 intronic, 103 exonic) to compute the number of SNP differences between
 175 all pairs of 1530 haplotypes derived from 765 wild-caught mosquitoes. We then used
 176 pairwise genetic distances to perform hierarchical clustering, and found that haplotypes
 177 carrying resistance alleles in codon 995 were grouped into 10 distinct clusters, each with
 178 near-identical haplotypes. Five of these clusters contained haplotypes carrying the L995F
 179 allele (labelled F1-F5), and a further five clusters contained haplotypes carrying L995S
 180 (labelled S1-S5).

181 To further investigate genetic backgrounds carrying resistance alleles, we used the
 182 Ag1000G haplotype data to construct median-joining networks [27] (Figure 2). The net-

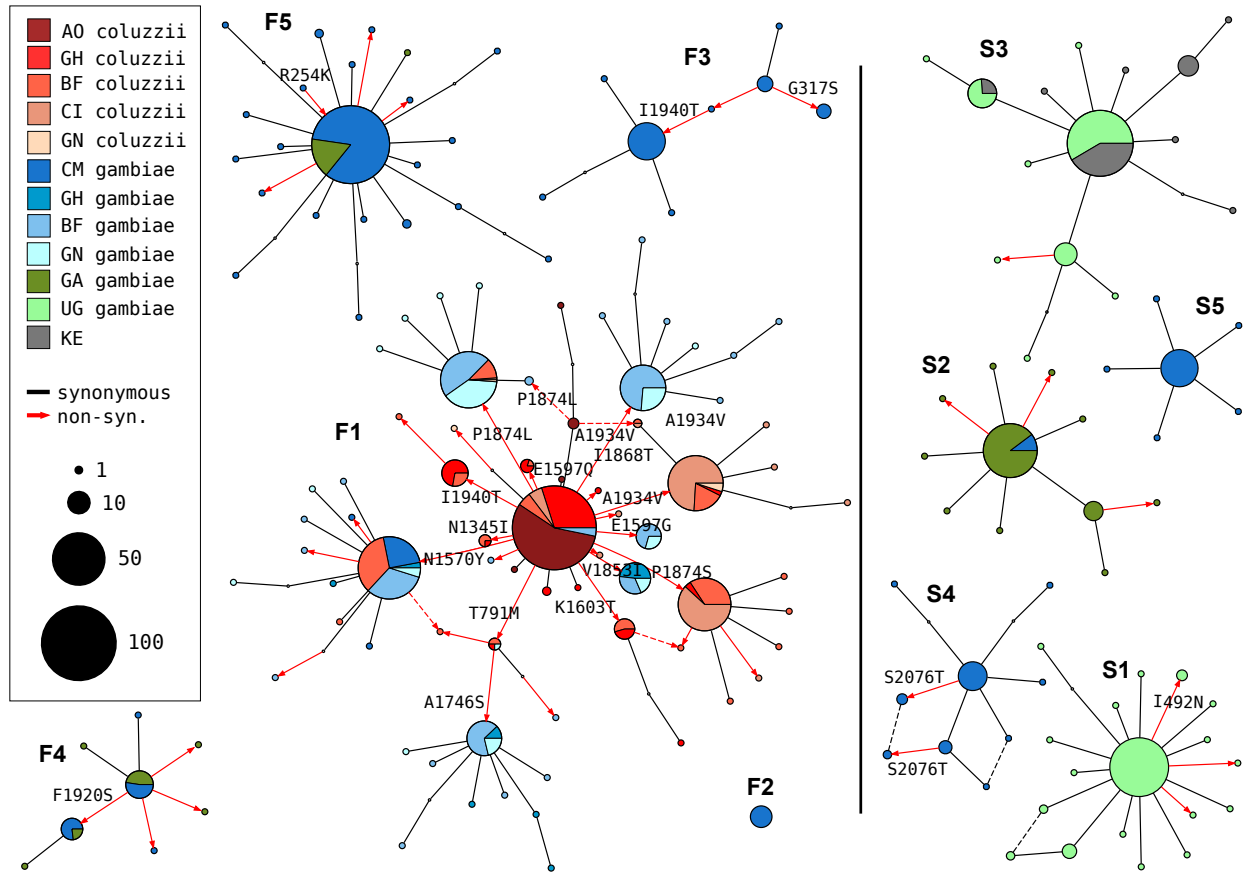


Figure 2. Haplotype networks. Median joining network for haplotypes carrying L995F (labelled F1-F5) or L995S variants (S1-S5) with a maximum edge distance of two SNPs. Labelling of network components is via concordance with hierarchical clusters discovered in [22]. Node size represents the proportion of haplotypes from mosquito populations/species - AO=Angola; BF=Burkina Faso; GN=Guinea; CM=Cameroon; GA=Gabon; UG=Uganda; KE=Kenya. Non-synonymous edges are highlighted in red and those leading to non-singleton nodes are labelled with the codon change, arrow head indicates direction of change away from the reference allele. Network components with fewer than three haplotypes are not shown.

work analysis improves on hierarchical clustering by allowing for the reconstruction and placement of intermediate haplotypes that may not be observed in the data. It also allows for non-hierarchical relationships between haplotypes, which may arise if recombination events have occurred between haplotypes. We constructed the network up to a maximum edge distance of 2 SNP differences, to ensure that each connected component captures a group of closely-related haplotypes. The resulting network contained 5 groups containing haplotypes carrying L995F, and a further 5 groups carrying L995S, in close correspondence with previous results from hierarchical clustering (96.8% overall concordance in assignment of haplotypes to groups).

192 The haplotype network brings into sharp relief the explosive radiation of amino acid sub-
 193 stitutions secondary to the L995F allele (Figure 2). Within the F1 group, nodes carrying
 194 non-synonymous variants radiate out from a central node carrying only L995F, suggest-
 195 ing that the central node represents the ancestral haplotype carrying L995F alone which
 196 initially came under selection, and these secondary variants have arisen subsequently as
 197 new mutations. Many of the nodes carrying secondary variants are large, consistent with
 198 positive selection and a functional role for these secondary variants as modifiers of the
 199 L995F resistance phenotype. The F1 network also allows us to infer multiple introgression
 200 events between the two species. The central (putatively ancestral) node contains hap-
 201 lotypes from individuals of both species, as do nodes carrying the N1570Y, P1874L and
 202 T791M variants. This structure is consistent with an initial introgression of the ancestral
 203 F1 haplotype, followed later by introgressions of haplotypes carrying secondary mutations.
 204 The haplotype network also illustrates the contrasting levels of non-synonymous varia-
 205 tion between L995F and L995S. Only two non-synonymous variants are present within the
 206 L995S groups, and both are at low frequency, thus may be neutral or mildly deleterious
 207 variants that are hitch-hiking on selective sweeps for the L995S allele.

208 The F1 group contained haplotypes from mosquitoes of both species, and from mosquitoes
 209 sampled in 4 different countries (Guinea, Burkina Faso, Cameroon, Angola) (Figure 3).
 210 The F4, F5 and S2 groups each contained haplotypes from both Cameroon and Gabon.
 211 The S3 group contained haplotypes from both Uganda and Kenya. The haplotypes
 212 within each of these groups were nearly identical across the entire span of the *Vgsc* gene
 213 ($\pi < 5.1 \times 10^{-5} bp^{-1}$). In contrast, diversity among wild-type haplotypes was two or-
 214 ders of magnitude greater (Cameroon *An. gambiae* $\pi = 1.4 \times 10^{-3} bp^{-1}$; Guinea-Bissau
 215 $\pi = 5.7 \times 10^{-3} bp^{-1}$). Thus it is reasonable to assume that each of these five groups con-
 216 tains descendants of an ancestral haplotype that carried a resistance allele and has risen in
 217 frequency due to selection for insecticide resistance. Given this assumption, these groups
 218 each provide evidence for adaptive gene flow between mosquito populations separated by
 219 considerable geographical distances.

220 A limitation of both the hierarchical clustering and network analyses is that they rely on
 221 genetic distances within a fixed genomic window from the start to the end of the *Vgsc* gene.
 222 *Anopheles* mosquitoes undergo homologous recombination during meiosis in both males

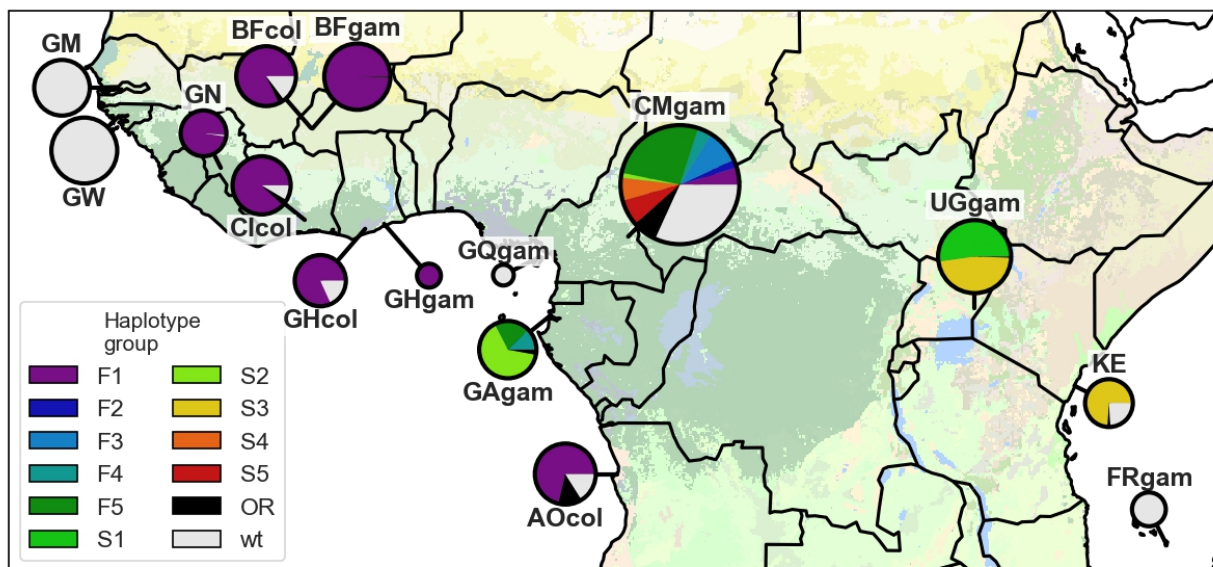


Figure 3. Map of haplotype frequencies. Each pie shows the frequency of different haplotype groups within one of the populations sampled. The size of the pie is proportional to the number of haplotypes sampled. The size of each wedge within the pie is proportional to the frequency of a haplotype group within the population. Haplotypes in groups F1-5 carry the L995F *kdr* allele. Haplotypes in groups S1-5 carry the L995S *kdr* allele. Haplotypes in group L1 carry the I1527T allele. Haplotypes in group L2 carry the M490I allele. Wild-type (*wt*) haplotypes do not carry any known or putative resistance alleles.

223 and females, and any recombination events that occurred within this genomic window
 224 could affect the way that haplotypes are grouped together in clusters or network compo-
 225 nents. In particular, recombination events could occur during the geographical spread of
 226 a resistance allele, altering the genetic background upstream and/or downstream of the
 227 allele itself. An analysis based on a fixed genomic window might then fail to infer gene flow
 228 between two mosquito populations, because haplotypes with and without a recombination
 229 event could be grouped separately, despite the fact that they share a recent common an-
 230 cestor. To investigate the possibility that recombination events may have affected our
 231 grouping of haplotypes carrying resistance alleles, we performed a windowed analysis of
 232 haplotype homozygosity, spanning *Vgsc* and up to a megabase upstream and downstream
 233 of the gene (Supplementary Figures S1, S2). This analysis supported a refinement of our
 234 initial grouping of haplotypes carrying resistance alleles. All haplotypes within groups S4
 235 and S5 were effectively identical on both the upstream and downstream flanks of the gene,
 236 but there was a region of divergence within the *Vgsc* gene itself that separated them in
 237 the fixed window analyses (Supplementary Figure S2). The 13.8 kbp region of divergence

occurred upstream of codon 995 and contained 8 SNPs that were fixed differences between S4 and S5. A possible explanation for this short region of divergence is that a gene conversion event has occurred within the gene, bringing a segment from a different genetic background onto the original genetic background on which the L995S resistance mutation occurred.

Positive selection for resistance alleles

To investigate evidence for positive selection on non-synonymous alleles, we performed an analysis of extended haplotype homozygosity (EHH) [28]. Haplotypes under recent positive selection will have increased rapidly in frequency, thus have had less time to be broken down by recombination, and should on average have longer regions of haplotype homozygosity relative to wild-type haplotypes. We defined a core region spanning *Vgsc* codon 995 and an additional 6 kbp of flanking sequence, which was the minimum required to differentiate the haplotype groups identified via clustering and network analyses. Within this core region, we found 18 distinct haplotypes at a frequency above 1% within the cohort. These included core haplotypes corresponding to each of the 10 haplotype groups carrying L995F or L995S alleles identified above, as well as a core haplotype carrying I1527T which we labelled L1 (due to it carrying the the wild-type leucine codon at position 995). We also found a core haplotype corresponding to a group of haplotypes from Kenya carrying an M490I allele, which we labelled as L2. All other core haplotypes we labelled as wild-type (*wt*). We then computed EHH decay for each core haplotype up to a megabase upstream and downstream of the core locus (Figure 4).

As expected, haplotypes carrying the L995F and L995S resistance alleles all experience a dramatically slower decay of EHH relative to wild-type haplotypes, supporting positive selection. Previous studies have found evidence for different rates of EHH decay between L995F and L995S haplotypes, suggesting differences in the timing and/or strength of selection [16]. However, we found no systematic difference in the length of shared haplotypes when comparing F1-5 (carrying L995F) against S1-5 (carrying L995S) (Supplementary Figure S3). There were, however, some differences between core haplotypes carrying the same allele. For example, shared haplotypes were significantly longer for S1 (median 1.091 cM, 95% CI [1.076 - 1.091]) versus other core haplotypes carrying L995S (e.g., S2 median

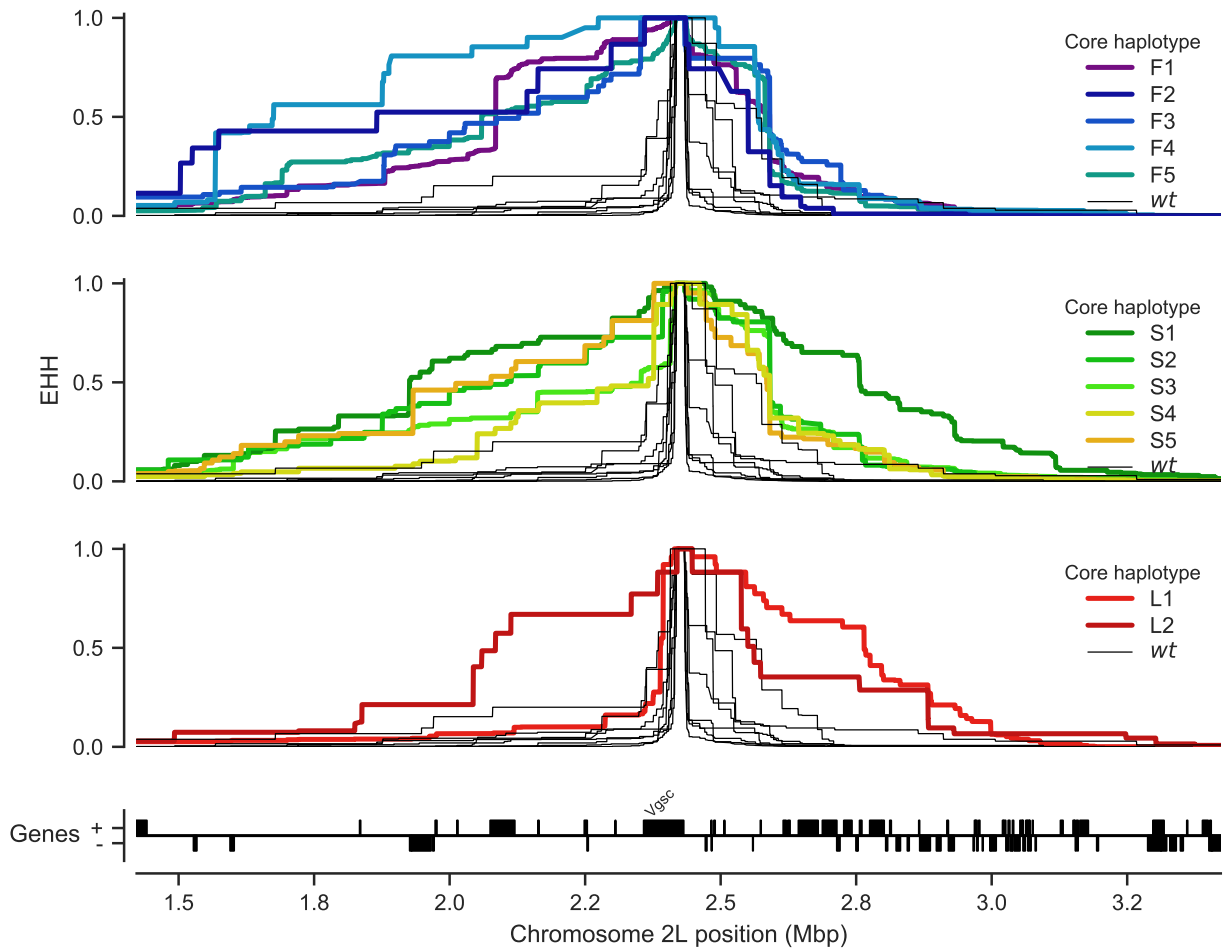


Figure 4. Evidence for positive selection on haplotypes carrying known or putative resistance alleles. Each panel plots the decay of extended haplotype homozygosity (EHH) for a set of core haplotypes centred on *Vgsc* codon 995. Core haplotypes F1-F5 carry the L995F allele; S1-S5 carry the L995S allele; L1 carries the I1527T allele; L2 carries the M490I allele. Wild-type (*wt*) haplotypes do not carry known or putative resistance alleles. A slower decay of EHH relative to wild-type haplotypes implies positive selection (each panel plots the same collection of wild-type haplotypes).

0.699 cM, 95% CI [0.696 - 0.705]; Supplementary Figure S3). Longer shared haplotypes indicate a more recent common ancestor, and thus some of these core haplotypes may have experienced more recent and/or more intense selection than others. The L1 haplotype carrying I1527T+V402L exhibited a slow decay of EHH on the downstream flank of the gene, similar to haplotypes carrying L995F and L995S, indicating that this combination of alleles has experienced positive selection. EHH decay on the upstream gene flank was faster, being similar to wild-type haplotypes, however there were two separate nucleotide substitutions encoding V402L within this group of haplotypes, and a faster EHH decay on this flank is consistent with recombination events bringing V402L alleles from differ-

ent genetic backgrounds together with an ancestral haplotype carrying I1527T. The L2 haplotype carrying M490I exhibited EHH decay on both flanks comparable to haplotypes carrying known resistance alleles. This could indicate evidence for selection on the M490I allele, however these haplotypes are derived from a Kenyan mosquito population where there is evidence for a severe recent bottleneck [22], and there were not enough wild-type haplotypes from Kenya with which to compare, thus this signal may also be due to the extreme demographic history of this population.

Discussion

Cross-resistance between pyrethroids and DDT

The VGSC protein is the physiological target of both pyrethroid insecticides and DDT [5]. The L995F and L995S alleles are known to increase resistance to both of these insecticide classes [7, 9]. By 2012, over half of African households owned at least one pyrethroid impregnated ITN and nearly two thirds of IRS programmes were using pyrethroids [2]. Pyrethroids were also introduced into agriculture in Africa prior to the scale-up of public health vector control programmes, and continue to be used on a variety of crops such as cotton [29]. DDT was used in Africa for several pilot IRS projects carried out during the first global campaign to eradicate malaria, during the 1950s and 1960s [12]. DDT is still approved for IRS use by WHO and remains in use in some locations, however within the last two decades pyrethroid use has been far more common and widespread. DDT was also used in agriculture from the 1940s, and although agricultural usage has greatly diminished since the 1970s, some usage remains [30]. In this study we reported evidence of positive selection on the L995F and L995S alleles, as well as the I1527T+V402L combination and possibly M490I. We also found 14 other non-synonymous substitutions that have arisen in association with L995F and appear to be positively selected. Given that pyrethroids have dominated public health insecticide use for two decades, it is reasonable to assume that the selection pressure on these alleles is primarily due to pyrethroids rather than DDT. It has previously been suggested that L995S may have been initially selected by DDT usage [16]. However, we did not find any systematic difference in the extent of haplotype homozygosity between these two alleles, suggesting that both alleles have been under selection over a

306 similar time frame. We did find some significant differences in haplotype homozygosity
 307 between different genetic backgrounds carrying resistance alleles, suggesting differences
 308 in the timing and/or strength of selection these may have experienced. However, there
 309 have been differences in the scale-up of pyrethroid-based interventions in different regions,
 310 and this could in turn generate heterogeneities in selection pressures. Nevertheless, it is
 311 possible that some if not all of the alleles we have reported provide some level of cross-
 312 resistance to DDT as well as pyrethroids, and we cannot exclude the possibility that
 313 earlier DDT usage may have contributed at least in part to their selection. The differing
 314 of resistance profiles to the two types of pyrethroids (type I, e.g., permethrin; and type
 315 II, e.g., deltamethrin) [31], will also affect the selection landscape. Further sampling and
 316 analysis is required to investigate the timing of different selection events and relate these
 317 to historical patterns of insecticide use in different regions.

318 **Resistance phenotypes for novel non-synonymous variants**

319 The sodium channel protein consists of four homologous domains (I-IV) each of which com-
 320 prises six transmembrane segments (S1-S6) connected by intracellular and extracellular
 321 loops [6]. Two pyrethroid binding sites have been predicted within the pore-forming mod-
 322 ules of the protein, the first (PyR1) involving residues from transmembrane segments IIS5
 323 and IIS6 and the internal linker between IIS4 and IIS5 (IIL45) [32], the second (PyR2)
 324 involving segments IS5, IS6, IIS6 and IL45 [25, 6]. Many of the amino acid substitutions
 325 known to cause pyrethroid resistance in insects affect residues within one of these two
 326 pyrethroid binding sites, and thus can directly alter pyrethroid binding [6]. For example,
 327 the L995F and L995S substitutions occur in segment IIS6 and belong to binding site PyR2
 328 [25]. The I1527T substitution that we discovered in *An. coluzzii* mosquitoes from Burk-
 329 ina Faso occurs in segment IIS6 and is immediately adjacent to two pyrethroid-sensing
 330 residues in site PyR1 [6]. It is thus plausible that pyrethroid binding could be altered by
 331 this substitution. The I1527T substitution (*M. domestica* codon 1532) has been found in
 332 *Aedes albopictus* [33], and substitutions in the nearby codon 1529 (*M. domestica* codon
 333 1534) have been reported in *Aedes albopictus* and in *Aedes aegypti* where it was found to be
 334 associated with pyrethroid resistance [6, 34, 35]. We found the I1527T allele in tight link-
 335 age with two alleles causing a V402L substitution (*M. domestica* codon 410). Substitutions

in codon 402 have been found in multiple insect species and are by themselves sufficient to confer pyrethroid resistance [6]. Codon 402 is within segment IS6, immediately adjacent to a pyrethroid sensing residue in site PyR2. The fact that we find I1527T and V402L in such tight mutual association is intriguing because (a) these two residues appear to affect different pyrethroid binding sites, and (b) haplotypes carrying V402L alone should also have been positively selected and thus be present in one or more populations.

A number of substitutions in segments of the protein that are not involved in either of the two pyrethroid binding sites have also been shown to confer pyrethroid resistance. For example, the N1570Y substitution causes substantially enhanced pyrethroid resistance when combined with L995F, although codon 1570 occurs in the internal linker between domains III and IV (LIII/IV) [25]. Computer modelling of the protein structure has suggested that substitutions in codon 1570 could allosterically alter site PyR2 and thus affect pyrethroid binding [25]. In addition to N1570Y, we found thirteen other substitutions at appreciable frequency occurring almost exclusively in association with L995F (Table 1; Figure 1). Of these, two (D466H, E1597G) occurred in the larger internal linkers between protein domains, one (R254K) occurred within a smaller internal linker between domain subunits, two (T791M, K1603T) occurred within an outer (“voltage-sensing”) transmembrane segment, one (A1746S) occurred within an inner (“pore-forming”) transmembrane segment, and the remaining seven occurred in the internal carboxyl-terminal tail. Thus there is no simple pattern regarding where these variants occur within the protein structure. Further work is required to confirm which of these substitutions affect pyrethroid resistance, and to determine whether they allosterically modify a pyrethroid binding site in a similar vein to N1570Y, or whether they provide some other benefit such as compensating for a deleterious effect of L995F on normal nervous system function. The novel M490I substitution, found on the Kenyan L2 haplotypic background potentially under selection, also occurs in an internal linker between protein domains (LI/II). However, M490I did not occur in association with L995F or any other non-synonymous substitutions. It is plausible that substitutions outside of pyrethroid binding sites could independently confer an insecticide resistance phenotype, because there are several known examples in other insect species [6]. Work in other species has also suggested that pyrethroid resistance substitutions could act not by altering pyrethroid binding but by altering the channel gating

kinetics or the voltage-dependence of activation [6]. Thus there are a number of potential mechanisms by which a pyrethroid resistance phenotype can be obtained, and clearly much remains to be unravelled regarding the molecular biology of pyrethroid resistance in this gene.

Design of genetic assays for surveillance of pyrethroid resistance

Entomological surveillance teams in Africa regularly genotype mosquitoes for resistance alleles in *Vgsc* codon 995, and use those results as an indicator for the presence of pyrethroid resistance alongside results from insecticide resistance bioassays. They typically do not, however, sequence the gene or genotype any other polymorphisms within the gene. Thus if there are other polymorphisms within the gene that cause or significantly enhance pyrethroid resistance, these will not be detected. Also, if a codon 995 resistance allele is observed, there is no way to know whether the allele is on a genetic background that has also been observed in other mosquito populations, and thus no way to investigate whether resistance alleles are emerging locally or being imported from elsewhere. Whole-genome sequencing of individual mosquitoes clearly provides data of sufficient resolution to answer these questions, and could be used to provide ongoing resistance surveillance. The cost of whole-genome sequencing continues to fall, with the present cost being approximately 50 GBP to obtain ~30× coverage of an individual *Anopheles* mosquito genome with 150 bp paired-end reads. However, to achieve substantial spatial and temporal coverage of mosquito populations, it is currently cheaper and more practical to develop targeted genetic assays for resistance outbreak surveillance. Technologies such as amplicon sequencing [36] could scale to tens of thousands of mosquitoes at low cost and could be implemented using existing platforms in national molecular biology facilities.

To facilitate the development of targeted genetic assays for surveillance of *Vgsc*-mediated pyrethroid resistance, we have produced several supplementary data tables. In Supplementary Table 1 we list all 64 non-synonymous variants found within the *Vgsc* gene in this study, with population allele frequencies. In Supplementary Table 2 we list 771 biallelic SNPs, within the *Vgsc* gene and up to 10 kbp upstream or downstream, that are potentially informative regarding which haplotype group a resistance haplotype belongs to, and thus could be used for tracking the spread of resistance. This table includes the allele

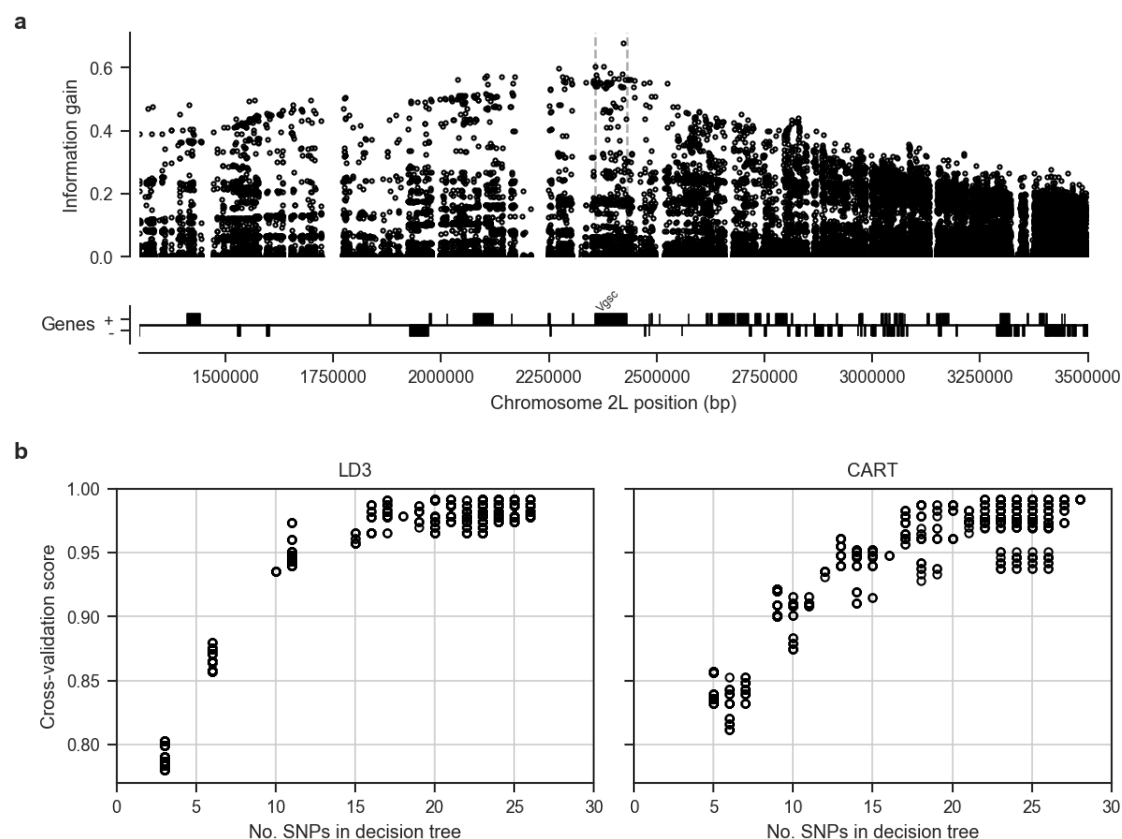


Figure 5. Informative SNPs for haplotype surveillance. **a**, Each data point represents a single SNP. The information gain value for each SNP provides an indication of how informative the SNP is likely to be if used as part of a genetic assay for testing whether a mosquito carries a resistance haplotype, and if so, which haplotype group it belongs to. **b**, Number of SNPs required to accurately predict which group a resistance haplotype belongs to. Each data point represents a single decision tree. Decision trees were constructed using either the LD3 (left) or CART (right) algorithm for comparison. Accuracy was evaluated using 10-fold stratified cross-validation.

frequency within each of the 12 haplotype groups defined here, to aid in identifying SNPs that are highly differentiated between two or more haplotype groups. We also provide Supplementary Table 3 which lists all 8,297 SNPs found within the *Vgsc* gene and up to 10 kbp upstream or downstream, which might need to be taken into account as flanking variation when searching for PCR primers to amplify a SNP of interest. To provide some indication for how many SNPs would need to be assayed in order to track the spread of resistance, we used haplotype data from this study to construct decision trees that could classify which of the 12 groups a given haplotype belongs to (Figure 5). This analysis suggested that it should be possible to construct a decision tree able to classify haplotypes with >95% accuracy by using 20 SNPs or less. In practice, more SNPs would be needed, to provide some redundancy, and also to type non-synonymous polymorphisms in

addition to identifying the genetic background. However, it is still likely to be well within the number of SNPs that could be assayed in a single multiplex via amplicon sequencing. Thus it should be feasible to produce low-cost, high-throughput genetic assays for tracking the spread of pyrethroid resistance. If combined with a limited amount of whole-genome sequencing at sentinel sites, this should also allow the identification of newly emerging resistance outbreaks.

Methods

Code

All scripts and Jupyter Notebooks used to generate analyses, figures and tables are available from the GitHub repository <https://github.com/malariagen/agam-vgsc-report>.

Data

We used variant calls from the Ag1000G Phase 1 AR3 data release (<https://www.malariagen.net/data/ag1000g-phase1-ar3>) and phased haplotype data from the Ag1000G Phase 1 AR3.1 data release (<https://www.malariagen.net/data/ag1000g-phase1-ar3.1>). Variant calls from Ag1000G Phase 1 are also available from the European Nucleotide Archive (ENA; <http://www.ebi.ac.uk/ena>) under study PRJEB18691.

Data collection and processing

For detailed information on Ag1000G WGS sample collection, sequencing, variant calling, quality control and phasing, see [22]. In brief, *An. gambiae* and *An. coluzzii* mosquitoes were collected from eight countries across Sub-Saharan Africa: Angola, Burkina Faso, Cameroon, Gabon, Guinea, Guinea Bissau, Kenya and Uganda. From Angola just *An. coluzzii* were sampled, Burkina Faso had samples of both *An. gambiae* and *An. coluzzii* and all other populations consisted of purely *An. gambiae*, except for Kenya and Guinea Bissau where species status is uncertain [22]. Mosquitoes were individually whole genome sequenced on the Illumina HiSeq 2000 platform, generating 100bp paired-end reads. Sequence reads were aligned to the *An. gambiae* AgamP3 reference genome assembly [37]. Aligned bam files underwent improvement, before variants were called using GATK Uni-

435 fiedGenotyper. Quality control included removal of samples with mean coverage $\leq 14\times$
436 and filtering of variants with attributes that were correlated with Mendelian error in ge-
437 netic crosses.

438 The Ag1000G variant data was functionally annotated using the SnpEff v4.1b soft-
439 ware [38]. Non-synonymous *Vgsc* variants were identified as all variants in transcript
440 AGAP004707-RA with a SnpEff annotation of “missense”. The *Vgsc* gene is known to
441 exhibit alternative splicing [5], however at the time of writing the *An. gambiae* gene an-
442 notations did not include the alternative transcripts reported by Davies et al. We wrote
443 a Python script to check for the presence of variants that are synonymous according to
444 transcript AGAP004707-RA but non-synonymous according to one of the other transcripts
445 present in the gene annotations or in the set reported by Davies et al. Supplementary Ta-
446 ble 1 includes the predicted effect for all SNPs that are non-synonymous in one or more
447 of these transcripts. None of the variants that are non-synonymous in a transcript other
448 than AGAP004707-RA were found to be above 5% frequency in any population.

449 For ease of comparison with previous work on *Vgsc*, pan Insecta, in Table 1 and Supple-
450 mentary Table 1 we report codon numbering for both *An. gambiae* and *Musca domestica*
451 (the species in which the gene was first discovered). The *M. domestica* *Vgsc* sequence
452 (EMBL accession X96668 [10]) was aligned with the *An. gambiae* AGAP004707-RA se-
453 quence (AgamP4.4 gene-set) using the Mega v7 software package [39]. A map of equiva-
454 lent codon numbers between the two species for the entire gene can be download from the
455 MalariaGEN website ([https://www.malariagen.net/sites/default/files/content/](https://www.malariagen.net/sites/default/files/content/blogs/domestica_gambiae_map.txt)
456 [blogs/domestica_gambiae_map.txt](https://www.malariagen.net/sites/default/files/content/blogs/domestica_gambiae_map.txt)).

457 Haplotypes for each chromosome of each sample were estimated (phased) using using
458 phase informative reads (PIRs) and SHAPEIT2 v2.r837 [40], see [22] supplementary text
459 for more details. The SHAPEIT2 algorithm is unable to phase multi-allelic positions,
460 therefore the two multi-allelic non-synonymous SNPs within the *Vgsc* gene, altering codons
461 V402 and M490, were phased onto the biallelic haplotype scaffold using MVNcall v1.0 [41].
462 Conservative filtering applied to the genome-wide callset had removed one of the three
463 known insecticide resistance conferring *kdr* variants, N1570Y [11]. Manual inspection of
464 the read alignment revealed that the SNP call could be confidently made, and it was
465 added back into the data set and then also phased onto the haplotypes using MVNcall.

466 Lewontin's D' [42] was used to compute the linkage disequilibrium (LD) between all pairs
467 of non-synonymous *Vgsc* mutations.

468 **Haplotype networks**

469 Haplotype networks were constructed using the median-joining algorithm [27] as imple-
470 mented in a Python module available from <https://github.com/malariagen/agam-vgsc-report>.
471 Haplotypes carrying either L995F or L995S mutations were analysed with a maximum edge
472 distance of two SNPs. Networks were rendered with the Graphviz library and a compos-
473 ite figure constructed using Inkscape. Non-synonymous edges were highlighted using the
474 SnpEff annotations [38].

475 **Positive selection**

476 Core haplotypes were defined on a 6,078 bp region spanning *Vgsc* codon 995, from chro-
477 mosome arm 2L position 2,420,443 and ending at position 2,426,521. This region was
478 chosen as it was the smallest region sufficient to differentiate between the ten genetic
479 backgrounds carrying either of the known resistance alleles L995F or L995S. Extended
480 haplotype homozygosity (EHH) was computed for all core haplotypes as described in
481 [28] using scikit-allel version 1.1.9 [43], excluding non-synonymous and singleton SNPs.
482 Analyses of haplotype homozygosity in moving windows (Supplementary Figs. S1, S2)
483 and pairwise haplotype sharing (Supplementary Figure S3) were performed using custom
484 Python code available from <https://github.com/malariagen/agam-vgsc-report>.

485 **Design of genetic assays for surveillance of pyrethroid resistance**

486 To explore the feasibility of indentifying a small subset of SNPs that would be sufficient
487 to identify each of the genetic backgrounds carrying known or putative resistance alleles,
488 we started with an input data set of all SNPs within the *Vgsc* gene or in the flanking
489 regions 20 kbp upstream and downstream of the gene. Each of the 1530 haplotypes in
490 the Ag1000G Phase 1 cohort was labelled according to which core haplotype it carried,
491 combining all core haplotypes not carrying known or putative resistance alleles together as
492 a single "wild-type" group. Decision tree classifiers were then constructed using scikit-learn
493 version 0.19.0 [44] for a range of maximum depths, repeating the tree construction process

10 times for each maximum depth with a different initial random state. The classification accuracy of each tree was evaluated using stratified 5-fold cross-validation.

References

- [1] S. Bhatt et al. ‘The effect of malaria control on *Plasmodium falciparum* in Africa between 2000 and 2015’. In: *Nature* 526.7572 (2015), pp. 207–211. ISSN: 0028-0836. arXiv: [arXiv:1011.1669v3](#).
- [2] Janet Hemingway et al. ‘Averting a malaria disaster: Will insecticide resistance derail malaria control?’ In: *The Lancet* 387.10029 (2016), pp. 1785–1788. ISSN: 1474547X.
- [3] World Health Organization. *Global Plan for Insecticide Resistance Management (GPIRM)*. Tech. rep. Geneva: World Health Organization, 2012.
- [4] World Health Organization et al. ‘Global vector control response 2017-2030.’ In: *Global vector control response 2017-2030*. (2017).
- [5] T. G.E. Davies et al. ‘A comparative study of voltage-gated sodium channels in the Insecta: Implications for pyrethroid resistance in Anopheline and other Neopteran species’. In: *Insect Molecular Biology* 16.3 (2007), pp. 361–375. ISSN: 09621075.
- [6] Ke Dong et al. ‘Molecular biology of insect sodium channels and pyrethroid resistance’. In: *Insect Biochemistry and Molecular Biology* 50.1 (2014), pp. 1–17. ISSN: 09651748.
- [7] D. Martinez-Torres et al. ‘Molecular characterization of pyrethroid knockdown resistance (kdr) in the major malaria vector *Anopheles gambiae* s.s.’ In: *Insect Molecular Biology* 7.2 (1998), pp. 179–184. ISSN: 09621075.
- [8] Ana Paula B Silva et al. ‘Mutations in the voltage-gated sodium channel gene of anophelines and their association with resistance to pyrethroids: a review’. In: *Parasites & Vectors* 7.1 (2014), p. 450. ISSN: 1756-3305.
- [9] H. Ranson et al. ‘Identification of a point mutation in the voltage-gated sodium channel gene of Kenyan *Anopheles gambiae* associated with resistance to DDT and pyrethroids’. In: *Insect Molecular Biology* 9.5 (2000), pp. 491–497. ISSN: 09621075.

- [10] Martin S. Williamson et al. ‘Identification of mutations in the housefly para-type sodium channel gene associated with knockdown resistance (kdr) to pyrethroid insecticides’. In: *Molecular and General Genetics* 252.1-2 (1996), pp. 51–60. ISSN: 00268925.
- [11] Christopher M Jones et al. ‘Footprints of positive selection associated with a mutation (N1575Y) in the voltage-gated sodium channel of *Anopheles gambiae*.’ In: *Proceedings of the National Academy of Sciences of the United States of America* 109.17 (2012), pp. 6614–9. ISSN: 1091-6490.
- [12] T. G. E. Davies et al. ‘DDT, pyrethrins, pyrethroids and insect sodium channels’. In: *IUBMB Life* 59.3 (2007), pp. 151–162. ISSN: 1521-6543.
- [13] Frank D. Rinkevich, Yuzhe Du and Ke Dong. ‘Diversity and convergence of sodium channel mutations involved in resistance to pyrethroids’. In: *Pesticide Biochemistry and Physiology* 106.3 (2013), pp. 93–100. ISSN: 00483575. arXiv: NIHMS150003.
- [14] J Pinto et al. ‘Multiple origins of knockdown resistance mutations in the Afrotropical mosquito vector *Anopheles gambiae*.’ In: *PLoS One* 2 (2007), e1243. ISSN: 19326203.
- [15] Josiane Etang et al. ‘Polymorphism of intron-1 in the voltage-gated sodium channel gene of *Anopheles gambiae* s.s. populations from cameroon with emphasis on insecticide knockdown resistance mutations’. In: *Molecular Ecology* 18.14 (2009), pp. 3076–3086. ISSN: 09621083.
- [16] Amy Lynd et al. ‘Field, genetic, and modeling approaches show strong positive selection acting upon an insecticide resistance mutation in *Anopheles gambiae* s.s.’ In: *Molecular Biology and Evolution* 27.5 (2010), pp. 1117–1125. ISSN: 07374038.
- [17] Federica Santolamazza et al. ‘Remarkable diversity of intron-1 of the para voltage-gated sodium channel gene in an *Anopheles gambiae*/*Anopheles coluzzii* hybrid zone.’ In: *Malaria journal* 14.1 (2015), p. 9. ISSN: 1475-2875.
- [18] Mylène Weill et al. ‘The kdr mutation occurs in the Mopti form of *Anopheles gambiae* s. through introgression’. In: *Insect molecular biology* 9.5 (2000), pp. 451–455.

- [19] Abdoulaye Diabaté et al. ‘The spread of the Leu-Phe kdr mutation through *Anopheles gambiae* complex in Burkina Faso: genetic introgression and de novo phenomena’. In: *Tropical Medicine & International Health* 9.12 (2004), pp. 1267–1273.
- [20] Chris S. Clarkson et al. ‘Adaptive introgression between *Anopheles* sibling species eliminates a major genomic island but not reproductive isolation’. In: *Nature Communications* 5 (2014). ISSN: 2041-1723.
- [21] Laura C. Norris et al. ‘Adaptive introgression in an African malaria mosquito coincident with the increased usage of insecticide-treated bed nets’. In: *Proceedings of the National Academy of Sciences* (2015), p. 201418892. ISSN: 0027-8424.
- [22] The *Anopheles gambiae* 1000 Genomes Consortium. ‘Natural diversity of the malaria vector *Anopheles gambiae*’. In: *Nature* 552 (2017), pp. 96–100.
- [23] L Wang et al. ‘A mutation in the intracellular loop III/IV of mosquito sodium channel synergizes the effect of mutations in helix IIS6 on pyrethroid resistance’. In: *Molecular Pharmacology* 87.3 (2015), pp. 421–429.
- [24] Shoji Sonoda et al. ‘Genomic organization of the para-sodium channel α -subunit genes from the pyrethroid-resistant and -susceptible strains of the diamondback moth’. In: *Archives of Insect Biochemistry and Physiology* 69.1 (2008), pp. 1–12. ISSN: 07394462.
- [25] Yuzhe Du et al. ‘Molecular evidence for dual pyrethroid-receptor sites on a mosquito sodium channel’. In: *Proceedings of the National Academy of Sciences* 110.29 (2013), pp. 11785–11790.
- [26] Kobié H. Toé et al. ‘Increased pyrethroid resistance in malaria vectors and decreased bed net effectiveness Burkina Faso’. In: *Emerging Infectious Diseases* 20.10 (2014), pp. 1691–1696. ISSN: 10806059.
- [27] H. J. Bandelt, P. Forster and A. Rohl. ‘Median-joining networks for inferring intraspecific phylogenies’. In: *Molecular Biology and Evolution* 16.1 (1999), pp. 37–48. ISSN: 0737-4038.
- [28] Pardis C. Sabeti et al. ‘Detecting recent positive selection in the human genome from haplotype structure’. In: *Nature* 419.6909 (2002), pp. 832–837. ISSN: 0028-0836.

- [29] Molly C Reid and F Ellis McKenzie. ‘The contribution of agricultural insecticide use to increasing insecticide resistance in African malaria vectors’. In: *Malaria journal* 15.1 (2016), p. 107.
- [30] Sara A Abuelmaali et al. ‘Impacts of agricultural practices on insecticide resistance in the malaria vector *Anopheles arabiensis* in Khartoum State, Sudan’. In: *PLoS One* 8.11 (2013), e80549.
- [31] Zhaonong Hu et al. ‘A sodium channel mutation identified in *Aedes aegypti* selectively reduces cockroach sodium channel sensitivity to type I, but not type II pyrethroids’. In: *Insect biochemistry and molecular biology* 41.1 (2011), pp. 9–13.
- [32] Andrias O. O’Reilly et al. ‘Modelling insecticide-binding sites in the voltage-gated sodium channel’. In: *Biochemical Journal* 396.2 (2006), pp. 255–263. ISSN: 0264-6021.
- [33] Jiabao Xu et al. ‘Multi-country survey revealed prevalent and novel F1534S mutation in voltage-gated sodium channel (VGSC) gene in *Aedes albopictus*’. In: *PLoS neglected tropical diseases* 10.5 (2016), e0004696.
- [34] Intan H Ishak et al. ‘Contrasting patterns of insecticide resistance and knockdown resistance (kdr) in the dengue vectors *Aedes aegypti* and *Aedes albopictus* from Malaysia’. In: *Parasites & vectors* 8.1 (2015), p. 181.
- [35] Yiji Li et al. ‘Evidence for multiple-insecticide resistance in urban *Aedes albopictus* populations in southern China’. In: *Parasites & vectors* 11.1 (2018), p. 4.
- [36] Andy Kilianski et al. ‘Bacterial and viral identification and differentiation by amplicon sequencing on the MinION nanopore sequencer.’ In: *GigaScience* 4 (2015), p. 12. ISSN: 2047-217X.
- [37] R A Holt et al. ‘The genome sequence of the malaria mosquito *Anopheles gambiae*’. In: *Science* 298.5591 (2002), pp. 129–149. ISSN: 0036-8075.
- [38] Pablo Cingolani et al. ‘A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3’. In: *Fly* 6.2 (2012), pp. 80–92. ISSN: 19336942.

- [39] Sudhir Kumar, Glen Stecher and Koichiro Tamura. ‘MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets’. In: *Molecular biology and evolution* 33.7 (2016), pp. 1870–1874. ISSN: 15371719.
- [40] Olivier Delaneau et al. ‘Haplotype estimation using sequencing reads’. In: *American Journal of Human Genetics* 93.4 (2013), pp. 687–696. ISSN: 00029297.
- [41] Androniki Menelaou and Jonathan Marchini. ‘Genotype calling and phasing using next-generation sequencing reads and a haplotype scaffold’. In: *Bioinformatics* 29.1 (2013), pp. 84–91. ISSN: 13674803.
- [42] R. C. Lewontin. ‘The Interaction of Selection and Linkage. I. General Considerations; Heterotic Models’. In: *Genetics* 49.1 (1964), pp. 49–67. ISSN: 0016-6731.
- [43] Alistair Miles and Nicholas Harding. *scikit-allel: A Python package for exploring and analysing genetic variation data*. 2016.
- [44] F. Pedregosa et al. ‘Scikit-learn: Machine Learning in Python’. In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

Supplementary figures

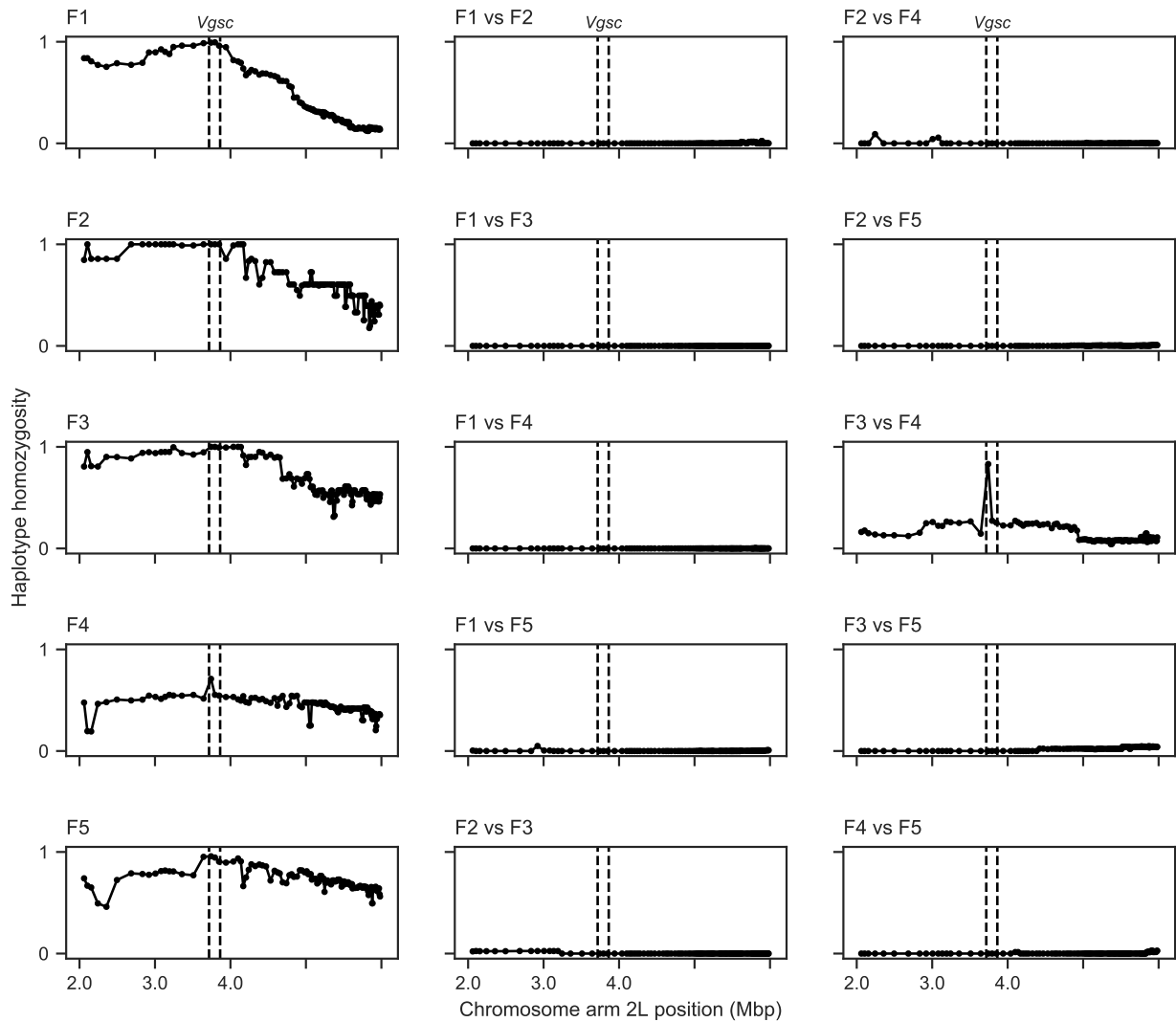


Figure S1. Windowed analysis of haplotype homozygosity for genetic backgrounds carrying the L995F allele. Each sub-plot shows the fraction of haplotype pairs that are identical within half-overlapping moving windows of 1000 SNPs. Each sub-plot in the left-hand column shows homozygosity for haplotype pairs within one of the haplotype groups identified by the network analysis. Sub-plots in the central and right-hand columns show homozygosity for haplotype pairs between two haplotype groups. If two haplotype groups are truly unrelated, haplotype homozygosity between them should be close to zero across the whole genome region. Dashed vertical lines show the location of the *Vgsc* gene.

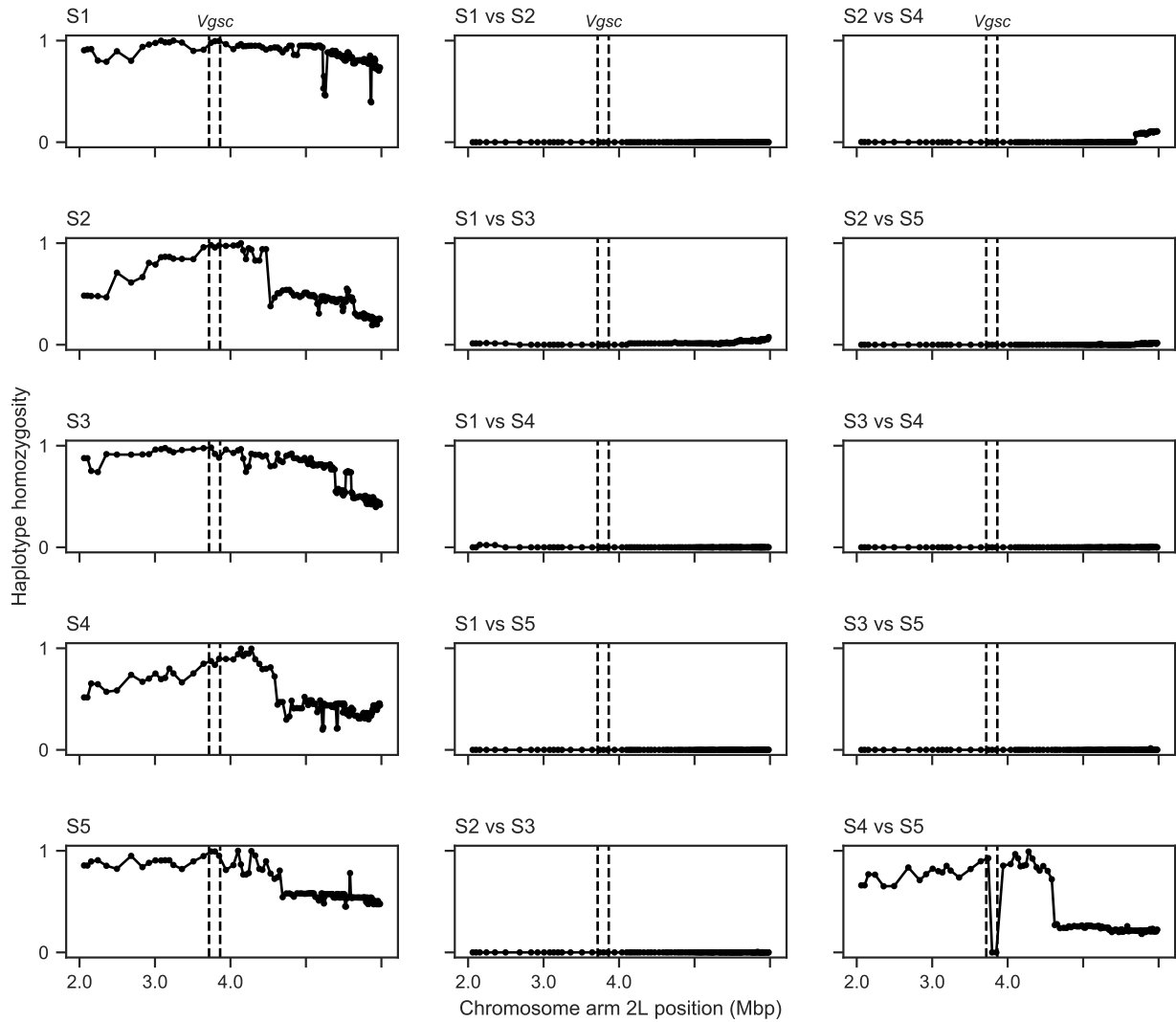


Figure S2. Windowed analysis of haplotype homozygosity for genetic backgrounds carrying the L995S allele. See Supplementary Figure S1 for explanation. Haplotype homozygosity is high between groups S4 and S5 on both flanks of the gene, indicating that haplotypes from both groups are in fact closely related.

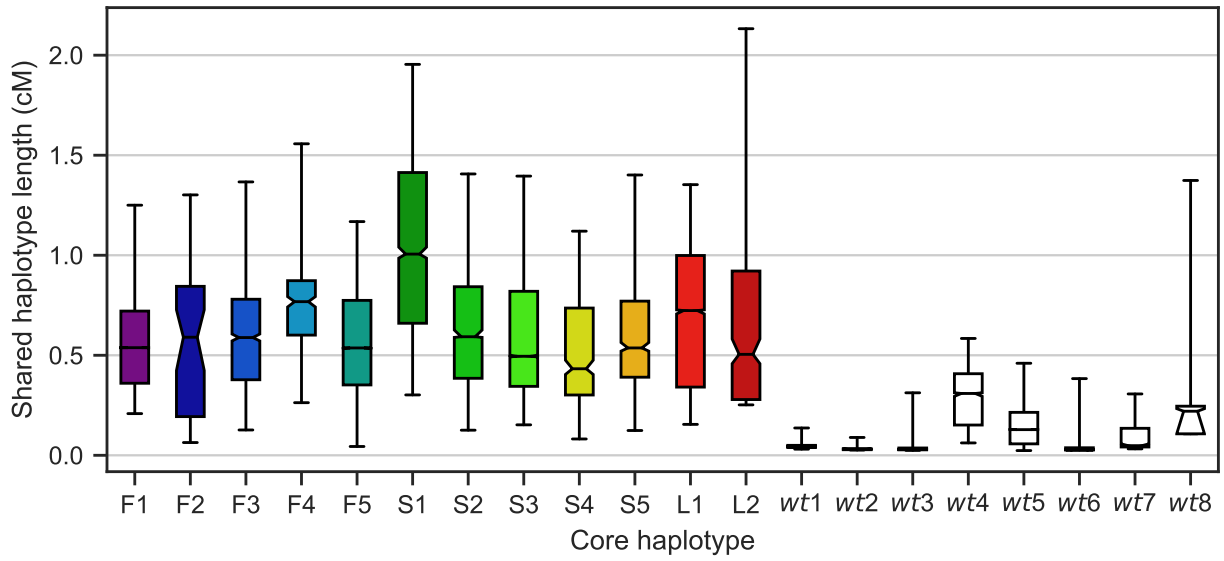


Figure S3. Shared haplotype length. Each bar shows the distribution of shared haplotype lengths between all pairs of haplotypes with the same core haplotype. For each pair of haplotypes, the shared haplotype length is computed as the region extending upstream and downstream from the core locus (*Vgsc* codon 995) over which haplotypes are identical at all non-singleton variants. The *Vgsc* gene sits on the border of pericentromeric heterochromatin and euchromatin, and we assume different recombination rates in upstream and downstream regions. The shared haplotype length is expressed in centiMorgans (cM) assuming a constant recombination rate of 2.0 cM/Mb on the downstream (euchromatin) flank and 0.6 cM/Mb on the upstream (heterochromatin) flank. Bars show the inter-quartile range, fliers show the 5-95th percentiles, horizontal black line shows the median, notch in bar shows the 95% bootstrap confidence interval for the median. Haplotypes F1-5 each carry the L995F resistance allele. Haplotypes S1-5 each carry the L995S resistance allele. Haplotype L1 carries the I1527T allele. Haplotype L2 carries the M490I allele. Wild-type (*wt*) haplotypes do not carry any known or putative resistance alleles.